Chi-squared: A simpler evaluation function for multiple-instance learning

Amy McGovern AMY@CS.UMASS.EDU
David Jensen JENSEN@CS.UMASS.EDU
Knowledge Discovery Lab, Computer Science Department, Univ. of Massachusetts Amherst, Amherst, MA 01003, USA

Abstract

This paper introduces a new evaluation function for solving the multiple instance problem. Our approach makes use of the main idea of diverse density (Maron, 1998; Maron & Lozano-Pérez, 1998) but finds the best concept using the chi-square statistic. This approach is simpler than diverse density and allows us to search more extensively by using properties of the contingency table to prune in a guaranteed manner. We demonstrate that this approach solves the multiple-instance problem as well as or better than diverse density and that the pruning mechanism allows chi-squared to identify the best concepts more quickly.

1. Introduction

Multiple instance learning (MIL) is a useful and well-known technique for learning with ambiguous or partially labeled data. For example, Dietterich et al. (1997) proposed the task of identifying what part of a molecule binds with a musk receptor in the nose. A molecule can take on a number of different shapes, or conformations, but it is likely that only one of those shapes will bind with a musk receptor. It is difficult to determine which shape actually matched the musk receptor although it is possible to observe the presence of a reaction for a given molecule (and thus a set of shapes). The task is to identify which shape and part of the molecules causes them to smell musky. The data available for a MIL agent is in the form of labeled sets of instances, e.g., a set of shapes with a single label. Supervised learning uses individually labeled instances which are difficult to obtain for many tasks. Labeling each instance in the set with the label for the set produces too much noise for supervised learning whereas MIL techniques are designed to learn from exactly this type of data.

More specifically, a MI learner uses labeled *bags* where a bag is a collection of *instances* with one label for the entire collection. A positive bag contains at least one instance of the *target concept* while a negative bag contains none. An instance is a point in feature space and the target concept is another point in feature space. The goal is to find a concept that explains the labels for the bags and can predict labels for unseen data. It is not known in advance which instance is the one causing the bag to be labeled as positive. If this were known, a supervised learning approach could be used instead.

Table 1. Contingency table used to identify the best concept from a set of positive and negative bags using the chi-squared statistic.

		Actual Bag label	
Predicted bag label		+	-
	+	a	b
	-	c	d

Successful applications of MIL include such tasks as recognizing MUSK molecules (Dietterich et al., 1997), predicting stock trends (Maron, 1998), image retrieval tasks (Goldman et al., 2002; Maron & Ratan, 1998), and identifying useful subgoals (McGovern & Barto, 2001a; McGovern, 2002). One of the main MI techniques is that of diverse density (Maron, 1998; Maron & Lozano-Pérez, 1998). This technique is widely used because of its intuitive appeal of its central idea: that the best concept is the concept closest to the intersection of the positive bags and farthest from the union of the negative bags. Because the nearby positive instances are from *different* bags, Maron denotes this idea as *diverse* density. Regular density algorithms would not take into account the different bags and would focus only at the instance level density. In spite of the intuitive appeal of diverse density, computing the diverse density evaluation function is computationally expensive.

We present an alternative evaluation function that uses the chi-square statistic. This is simpler to define and implement and it allows a guaranteed pruning method. In particular, we use the contingency table shown in Table 1. The

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2003		**00-00-2003 to 00-00-2003**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Chi-squared: A simpler evaluation function for multiple-instance learning	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Massachusetts Amherst,Knowledge Discovery Laboratory,140 Governors Drive,Amherst,MA,01003**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**8**	

rows of the table correspond to the predicted label for the bag and the columns of the table correspond to the labels on the training bags. Using the molecule example, if the current hypothesis predicts two of three positive bags as positive and three of four negative bags as negative, then the contingency table could be filled out as:

2	1
1	3

The concept with a maximal chi-squared value will have most of its mass concentrated along the main diagonal and thus will be correctly predicting the most positive and negative bags. This correlates with the main idea of diverse density by correctly predicting as many positive bags as possible while also predicting negative bags. The use of the contingency table of this form enables us to also introduce a guaranteed pruning method similar to (Oates & Cohen, 1996; Webb, 1995). We discuss the pruning method in more detail below.

2. Notation

For the MI notation, we follow that of Maron (1998) and Maron and Lozano-Pérez (1998). The set of positive bags is denoted B^+ and the ith positive bag is B_i^+. Likewise, the set of negative bags is denoted B^- and the ith negative bag is B_i^-. If the discussion applies to both types of bags, we drop the superscript and refer to it as B. The jth instance of the ith bag is denoted B_{ij}. The target concept is denoted c_t and other concepts as c.

3. Diverse density

We first review diverse density and then examine how our method relates. We use diverse density as the baseline for comparison in the experimental results. Maron and Lozano-Pérez (1998) and Maron (1998) define the diverse density of a concept, c, to be:

$$DD(c) = Pr(c|B_1^+,\ldots,B_n^+,B_1^-,\ldots,B_m^-), \quad (1)$$

where $Pr(c)$ is the probability that c is the correct concept, n is the number of positive bags, and m is the number of negative bags. The output of a DD search is the concept with a maximal DD value. To perform this search, we must expand Equation 1. Using Bayes' Rule, Equation 1 can be rewritten as:

$$DD(c) = \frac{Pr(B_1^+,\ldots,B_n^+,B_1^-,\ldots,B_m^-|c)Pr(c)}{Pr(B_1^+,\ldots,B_n^+,B_1^-,\ldots,B_m^-)}. \quad (2)$$

Assuming a uniform prior probability over the target concepts and noticing that the denominator is constant with respect to the concept, finding the concept with the maximum DD value reduces to finding the maximum likelihood of the positive and negative bags given a specific concept:

$$Pr(B_1^+,\ldots,B_n^+,B_1^-,\ldots,B_m^-|c).$$

Maron assumes that the bags are conditionally independent given the target concept, which allows the likelihood to be rewritten as:

$$\prod_{1\le i\le n} Pr(B_i^+|c) \prod_{1\le i\le m} Pr(B_i^-|c). \quad (3)$$

However, it is still not possible to calculate this exactly without a model of how the bags were generated. Instead, one can use Bayes' Rule again to rewrite Expression 3 as:

$$\prod_{1\le i\le n} \frac{Pr(c|B_i^+)Pr(B_i^+)}{Pr(c)} \prod_{1\le i\le m} \frac{Pr(c|B_i^-)Pr(B_i^-)}{Pr(c)}. \quad (4)$$

Substituting Expression 4 into Equation 2 and omitting the terms that do not depend on c, the task reduces to that of finding the maximum likelihood over concepts as follows:

$$\prod_{1\le i\le n} Pr(c|B_i^+) \prod_{1\le i\le m} Pr(c|B_i^-).$$

It remains to determine the probability of an instance in a bag causing the concept to be correct, $Pr(c|B_i)$. Maron discusses several ways to do this. We follow his suggestion of using a noisy-or model (Pearl, 1988), in which case we have:

$$Pr(c|B_i^+) = 1 - \prod_{1\le j\le p} (1 - Pr(B_{ij}^+ \in c)), \text{ and} \quad (5)$$

$$Pr(c|B_i^-) = \prod_{1\le j\le p} (1 - Pr(B_{ij}^- \in c)), \quad (6)$$

where p is the number of instances in bag B_i.

The only part of Equations 5 and 6 that is undefined is the probability of a particular instance belonging to the target concept: $Pr(B_{ij} \in c)$. This can be defined in several ways. For the results presented here, we use Maron's single point concept class which he defines using a Gaussian probability distribution. The concept, c, is assumed to be a k dimensional vector and the probability of an instance belong to the concept class is:

$$Pr(B_{ij} \in c) = \frac{\exp\left(\frac{-\sum_{1\le l\le k}\left(B_{ijl}-c_l\right)^2}{\sigma^2}\right)}{Z}, \quad (7)$$

where B_{ijl} is the lth feature of the jth instance of the ith bag, c_l is the lth feature of concept c, Z is a scaling factor, and σ^2 is the standard deviation. The standard deviations σ_+^2 and σ_-^2 can be chosen separately for positive and negative bags to allow the two types of evidence to have different amounts of influence on the DD values.

Note that, in practice, it is more accurate to calculate the log likelihood of the diverse density, rather than calculating the diverse density directly, because of accuracy issues with very small floating point numbers.

3.1. Identifying a concept with maximal DD

Concepts with maximal DD values can be found using standard search techniques. If it is feasible, e.g., if the space is small and can be viewed as a discrete space, exhaustive search can be used. In larger spaces, Maron used a gradient based search approach with multiple restarts where each run of the gradient search started from a random point in a positive bag. This heuristic is useful because the MI framework already restricts the true concept to appear only in the positive bags. However, gradient methods have several drawbacks. The first is that, even with the heuristics on starting locations, gradient methods can become stuck in local maxima. Second, they require the function being maximized to be differentiable. While this is true for the concepts that Maron proposed such as the single point concept and for the linear concept class (McGovern & Barto, 2001b), it is not true for other types of concepts such as relational graphs (McGovern & Jensen, 2003).

One alternative to the gradient search for diverse density was proposed by Zhang and Goldman (2002). They combined the expectation-maximization framework with the search for a concept with maximal DD value. This approach yields high performance but is again computationally expensive. Another alternative in large spaces is a simple random search technique (Rosenstein & Barto, 2001). This is the search method that we use for the experiments presented in this paper where exhaustive search is not feasible. In this case, the search starts from a number of random locations such as randomly chosen instances from the positive bags, and the search proceeds in a manner similar to genetic algorithms. The random search method can still become stuck in local maxima but it can be used on functions that are not differentiable.

4. Chi-squared

We propose an alternative way to identify the best concept in an MI setting that is based on the main idea of diverse density but relies on the chi-squared statistic. This is a well known statistic with a known sampling distribution. The contingency table used to calculate chi-squared can also be used to prune the search.

The main idea of diverse density is that the best concept is the concept that is at the intersection of the positive bags and that is far away from the union of the negative bags. Using the contingency table shown in Table 1, chi-squared identifies the concepts that predict the most positive bags as well as the most negative bags. This is related to the most diversely dense concept although it does not use the idea of being far away from the negative bags. The rows of the table correspond to the predicted label from the concept and the columns correspond to the actual labels for the training bags. Assuming a method for labeling the bags given a proposed target concept, the table is filled out in the following manner. If the concept predicts that the bag will be positive and it is positive, a is incremented. If the prediction is positive but the bag is really negative, b is incremented. If the prediction is negative and the bag is positive, c is incremented. Finally, d is incremented if the concept predicts negative and the bag is negative.

Chi-squared is calculated by summing the squared differences for the expected values in each cell of the contingency table versus the observed values. Let o_a, o_b, o_c, and o_d be the observed values for the cells of the table shown in Table 1 and let e_a, e_b, e_c, and e_d be the expected values for each of the cells. The expected values are calculated by multiplying the row sum by the column sum and dividing by the total number of elements. The chi-squared statistic is calculated as:

$$\chi^2 = \sum_{i \in a,b,c,d} \frac{(o_i - e_i)^2}{e_i}$$

The best concept is defined as that with the highest chi-squared value. Chi-squared will be maximal in two cases: when the mass is concentrated along the main diagonal (e.g., in a and d) and when the mass is concentrated along the off-diagonal (e.g., in b and c). In the first case, the proposed concept is correctly predicting a maximum number of positive and negative bags, which is the overall goal. In the second case, the concept is predicting exactly the opposite of this goal. This is a well-known issue with the chi-squared statistic and the signed chi-squared statistic addresses this issue. We define the best concept to have a maximal signed chi-squared value. Signed chi-squared is positive if the mass is on the main diagonal and negative if it is on the off-diagonal.

There are several advantages to calculating the chi-squared statistic over calculating diverse density. The first is it is both simpler to calculate as well as less computationally complex for search. Second, this approach can be used for concept spaces that are not differentiable or are not able to be defined clearly in the diverse density framework. Maron's approach assumes that the DD equation is differentiable which means that the probability $Pr(B_{ij} \in c)$ must be differentiable. Although this is generally true in a flat feature space, it is not true for relational data yet we can successfully apply the chi-squared technique to such tasks (McGovern & Jensen, 2003). A third, and very important, advantage of the chi-squared technique over other MI techniques is that the chi-squared approach enables a guaranteed pruning mechanism. This means that the signed chi-squared approach should be more effective in larger spaces or with a limited amount of time to search because it can search more thoroughly in the same amount of time.

4.1. Pruning

Any search technique that can make use of pruning can be used to identify the best chi-squared concept. Pruning works in a manner very similar to that of Oates and Cohen (1996) and Webb (1995). In particular, assume a concept x is being proposed as a target concept and that the contingency table for this concept has values:

a	b
c	d

. This contingency table can be evaluated to give a chi-squared value but it can also be used to find the maximum possible chi-squared value for a concept based on x but that is more specific than x. A more specific concept could be one where σ_+^2 or σ_-^2 is smaller or, for a graphical concept, one where the graph has more nodes or edges.

A more specific concept is unable to match *more* bags than the original concept, so the mass in the contingency table is restricted to move from the top row (i.e. positive predictions) to the bottom row. Also, since the columns of the table are the labels on the bags in the training set, the mass can not move from one column to another without the training data being relabeled. With these restrictions, a concept based on concept x would have a maximal chi-squared value at either

0	b
a + c	d

or

a	0
c	b + d

. In the first case, the signed chi-squared value will actually be minimized and the concept can be immediately pruned as it is predicting exactly the opposite of the correct bag labels. In the second case, if the maximum possible value that a concept based on x can reach is worse than the best concept seen so far in the search, then x can be pruned and the search moved to a new area.

4.2. Labeling the bags

The cells in the contingency table are filled by predicting the label on each of the training bags using the proposed target concept. Using the DD definitions, there are several ways to do this that could make sense. The first is to calculate the probability that an instance is in the concept, $Pr(B_{ij} \in c)$, for each instance in bag B_i and to take the maximum value, e.g.,

$$\text{label}_{\max}(B_i|c) = \max_{1 \le j \le p} Pr(B_{ij} \in c)$$

In this case, a label of one is a prediction of a positive bag and a label of zero is a prediction of a negative bag. A related approach could use the sum of these probabilities:

$$\text{label}_{\text{add}}(B_i|c) = \sum_{1 \le j \le p} Pr(B_{ij} \in c)$$

Other approaches that might seem useful that are based on the product or the sum of the logarithms of $Pr(B_{ij} \in c)$ (similar to the calculations that Maron defines for $Pr(c|B_i)$) do not work well in practice.

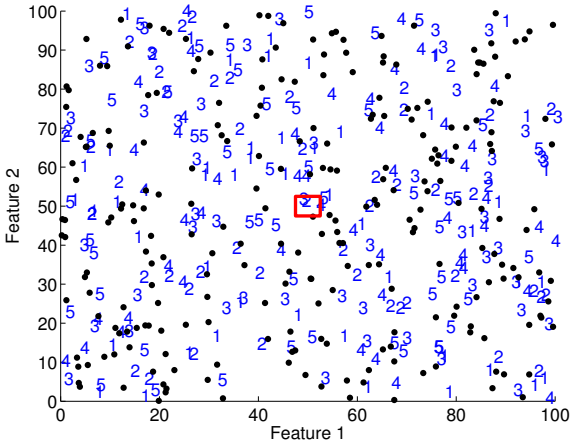

Figure 1. Sample artificial data set where there are five positive bags and five negative bags. Positive instances are labeled with a number corresponding to the bag and negative instances are shown as dots. The rectangle shows the target concept. This data set is derived from (Maron, 1998).

Using the $\text{label}_{\max}$ approach, the cells of the table are filled out as follows. Given a proposed concept x and a positive bag B_i^+, a is incremented with $\text{label}_{\max}(B_i^+, x)$ and c by $1 - \text{label}_{\max}(B_i^+, x)$. For negative bags, b is incremented with $\text{label}_{\max}(B_i^-, x)$ and d with $1 - \text{label}_{\max}(B_i^-, x)$. For a set of bags B^+ and B^- and a proposed concept x, the contingency table becomes:

$\sum_{i=1}^{m} \text{label}_{\max}(B_i^+, x)$	$\sum_{i=1}^{n} \text{label}_{\max}(B_i^-, x)$
$\sum_{i=1}^{m} 1 - \text{label}_{\max}(B_i^+, x)$	$\sum_{i=1}^{n} 1 - \text{label}_{\max}(B_i^-, x)$

The table for $\text{label}_{\text{add}}$ is constructed in a similar manner.

Under any labeling technique of this form, when the concept correctly predicts the most positive and negative bags, the mass in the contingency table will be concentrated along the main diagonal. This means that the concept will have a maximal chi-squared value. Each incorrect prediction will decrease the signed chi-squared value by adding mass off diagonal.

In the next section, we compare diverse density and the signed chi-squared approaches more directly.

5. Experimental results

We compare the behavior of the diverse density and signed chi-squared methods under varying conditions using Maron's 'difficult artificial data set' (Maron, 1998). We use this data set because it allows us to control many aspects of the data available to the two MI learners while still presenting a challenging task. The feature space is two dimensional and each feature is a real number in the range $[0, 100]$. The target concept is a 5x5 square in the middle of the space surrounding the point (50,50). Bags are created

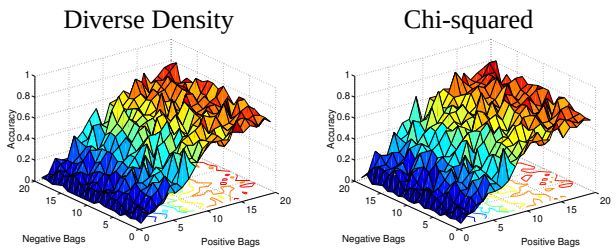

Figure 2. Accuracy of the diverse density and signed chi-squared methods with a varying number of positive and negative bags. These numbers are averaged over 60 different runs.

by uniformly sampling 50 points from the space and adding these instances to each bag. If any point falls inside the target concept, then the bag is labeled as positive, otherwise the bag is labeled as negative.

Figure 1 shows an example data set with five positive bags and five negative bags. This task is difficult not only because of the type of data available but also because of the target concept itself. The concept that the MI learner is searching for is a point in feature space while the actual shape of the concept is a square. The single-point concept class surrounds each point with a Gaussian-like sphere but this means that points inside a square may be missed, depending on the radius of the sphere. This makes the task more challenging than if we randomly generated target points and used a sphere around them to define the target concept.

For the first experiments, we repeat Maron's measurements of accuracy using diverse density with a varying number of positive and negative bags. We also measure accuracy using the chi-squared method under the same conditions. In this case, we discretized the feature space using unit-sized grid squares and computed the negative log-likelihood diverse density values and the chi-squared values for each point in the discretized space. We then identified the best concepts for both methods. Accuracy is measured by checking if the best concept fell within the target region. If this was the case, that run received a score of one and zero otherwise. We measured the average accuracy over 60 different runs, each starting with a different random seed to create the bags. We varied the number of positive bags from three to twenty and the number of negative bags from one to twenty. The average accuracy results are shown in Figure 2. The chi-squared results shown in this figure used the $\text{label}_{\text{max}}$ approach. The two methods have comparable accuracies for the same number of positive and negative bags, which is expected.

It is difficult to tell from these graphs if either approach is outperforming the other by a small but significant amount. To determine this, we average over the varying number

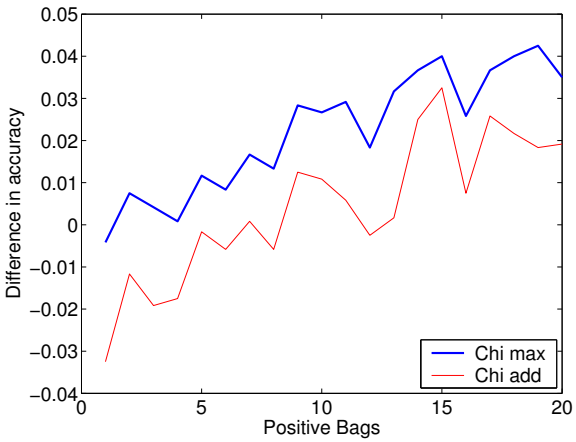

Figure 3. Average difference in accuracy for a varying number of positive bags (averaged over the varying number of negative bags) for the chi-squared approach using the $\text{label}_{\text{max}}$ and $\text{label}_{\text{add}}$ to label the bags.

of negative bags and examine the difference between the methods as a function of the number of positive bags. Figure 3 shows the average accuracy of chi-squared minus the accuracy of diverse density using both the $\text{label}_{\text{max}}$ and $\text{label}_{\text{add}}$ approaches. There are two things to note from this picture. First, the $\text{label}_{\text{max}}$ approach completely dominates the $\text{label}_{\text{add}}$ approach. Knowing this, we only present results using the $\text{label}_{\text{max}}$ approach for the rest of this paper. Second, the chi-squared approach using $\text{label}_{\text{max}}$ outperforms diverse density by a few percent with this difference peaking at around four percent better than diverse density.

5.1. Parameter variations

One of the main hypotheses of this work is that the chi-squared method will perform comparably or better than diverse density under a variety of conditions. With this in mind, we varied several parameters and conditions of the experiment and compared the accuracy of the two algorithms.

The first parameter that we varied was the number of instances in each bag. The results reported so far used the parameters from Maron (1998) as we wanted to duplicate these results for diverse density. In this experiment, we fixed the number of positive bags at 20 and the number of negative bags at 20 and varied the number of instances in each bag from 5 to 750. The reason that the number of instances in each bag could make a difference is that as the bag size decreases to one, the problem shifts from an MI task to a supervised learning task because the inherent noise of the task decreases (fewer instances per bag means that the true positive instances are easier to identify).

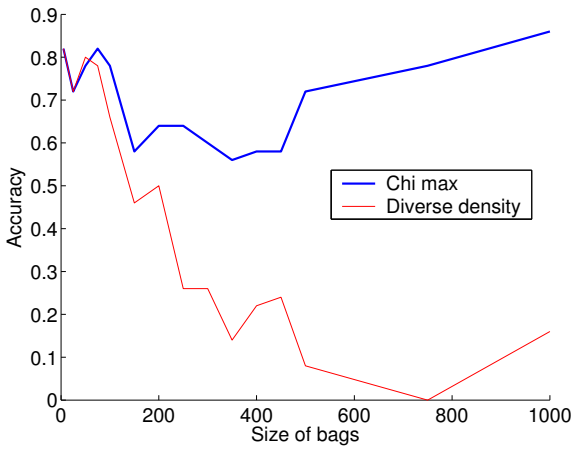

Figure 4. Average accuracy for diverse density and chi-squared with a varying number of instances in each bag. There were 20 positive bags and 20 negative bags.

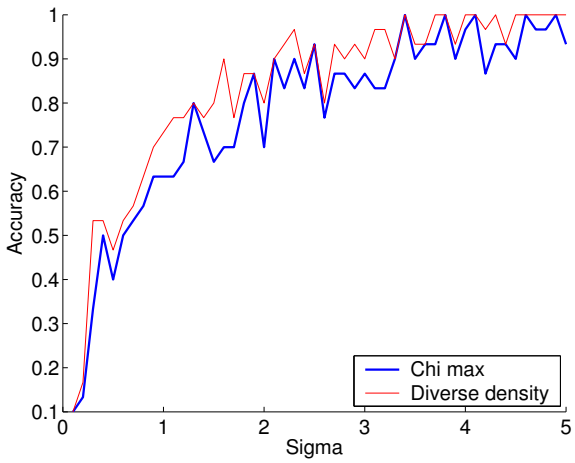

Figure 5. Average accuracy for diverse density and chi-squared when σ varies from 0.1 to 5.0.

A comparison of the accuracy of the two algorithms is shown in Figure 4. These numbers are averaged over 50 different runs, where each run has different instances in the bags. For bags with up to 100 instances per bag, there is no difference between the two methods. However, as the bags continue to grow in size, the signed chi-squared method significantly outperforms diverse density. To obtain these results, we used exhaustive search which means that the differences are not due to differences in search ability and depend only on the definition of the best concept for each method. With a constant number of bags, the problem difficulty increases as the number of instances in the bag increases. These results indicate that the chi-squared approach scales better with this aspect of task difficulty. We hypothesize that this is due to the $\text{label}_{\text{max}}$ approach which identifies the maximum match from a bag. This should minimize the effect of a bag's size on the algorithms ability to identify the best concept. Diverse density instead multiplies a number of individual probability calculations together over a bag which means that the larger bags may obscure the data more.

As discussed above, one reason that this data set presents a difficult task is that the size of the Gaussian ball around around the target concept was smaller than the target square. This can be adjusted by varying σ_-^2 and σ_+^2. We expected the two algorithms to perform comparably with this parameter variation because both relied on the same underlying calculation of $Pr(B_{ij} \in c)$. Figure 5 shows the results of varying $\sigma_+^2 = \sigma_-^2$ from 0.1 to 5.0 in increments of 0.1. As expected, the performance of the two algorithms was comparable.

5.2. The power of pruning

One of the main advantages of the chi-squared approach over diverse density is that it offers a guaranteed pruning method. We expected that this would give a significant advantage to the chi-squared method as the task was increased in difficulty. To examine this hypothesis, we used diverse density and chi-squared on several tasks of varying difficulty. In the first task, we varied the dimensions of the feature vector used to describe the instance and the concept. As the number of dimensions was increased, the size of the potential concept space increased exponentially. This means that pruning should be more useful as the size of the concept space increases.

One difficulty with varying only the number of features is that this approach assumes that the same proportion of data is available for each new dimension. This can be accomplished in two ways: by increasing the number of bags or by increasing the size of the bags. If we increase the number of bags, the total number of instances may remain in constant proportion over the varying dimensions but the task is actually made easier by having more positive and negative bags available as evidence. This is illustrated in Figure 6 where we compare the average accuracy of diverse density and chi-squared when each uses a number of search steps based on the dimensionality of the task. We used the simple random search method for both techniques[1]. In this case, the chi-squared approach significantly outperformed diverse density at the lower dimensions but as the number of bags increased with the size of the problem, both algorithms began to improve their accuracies.

A second way to vary the amount of data available to the

[1]Both methods relied on the same underlying code so as to ensure no difference in performance due to coding.

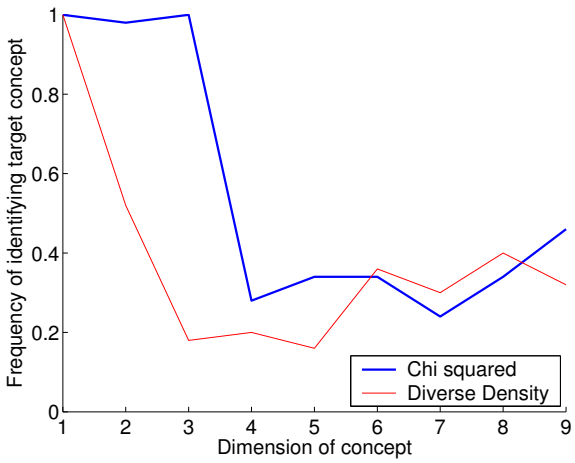

Figure 6. Average accuracy for diverse density and chi-squared with a fixed number of steps allowed for the search. The dimensionality of the feature space varies from 1 to 9 and the number of bags varied as a function of the number of dimensions.

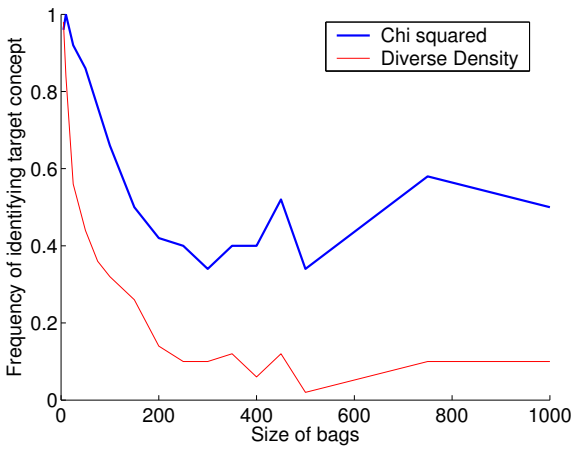

Figure 7. Average accuracy for diverse density and chi-squared using search in a two dimensional space where the number of instances in each bag varied from 5 to 1000.

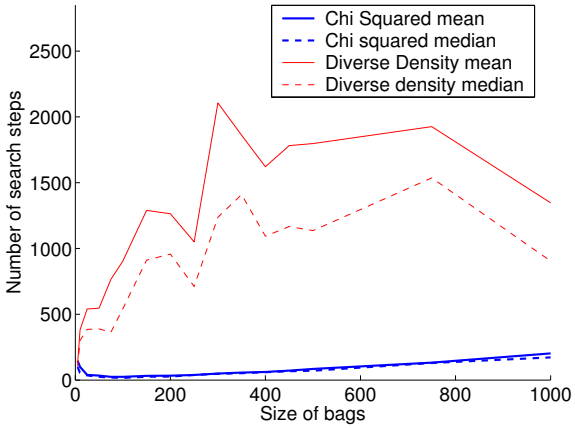

Figure 8. Average and median number of search steps used to identify the best concept with diverse density and chi-squared in a two dimensional space where the number of instances in each bag varied from 5 to 1000.

MI learner is to keep the number of bags constant but to increase the number of instances per bag as a function of the size of the problem space. This should significantly increase the difficulty of the problem as both the size of the space increases and the ability to identify the true positive instances from the positive bags decreases. We compared the accuracy of the two evaluation functions under this condition and we measured the total number of steps used for search in the two cases. In this case, the chi-squared approach is able to achieve higher accuracy than diverse density in fewer steps. As the difficulty of the problem is increased in both dimension and number of instances, both methods begin to drop in accuracy as well as to take more total steps to find the best concept.

We also compared the accuracy of the two methods with two features but with a varying number of instances per bag. This corresponds to the results presented earlier (c.f., Figure 4) but we used search with pruning instead of exhaustive search over the entire space. These results are presented in Figure 7. In this case, there were 20 positive bags and 20 negative bags with the number of instances in each bag varying from 5 to 1000. The accuracy of the diverse density approach drops very quickly as the size of the bags is increased while the accuracy of chi-squared drops more slowly and levels off at a better value.

Figure 8 shows the number of search steps used to identify the best concept for both diverse density and chi-squared in this experiment. Chi-square takes almost a constant number of steps to find the best concept while the number of steps required by diverse density grows in an polynomial manner. This is significant because it highlights one of the main advantages of the chi-squared approach: the ability to prune the search effectively and accurately.

6. Discussion and Conclusions

In summary, the chi-squared method that we introduced is a simpler approach than diverse density and it relies on a well known statistic. Using chi-squared allows a guaranteed pruning mechanism which significantly increases the efficiency of the search for the best concept. We demonstrated that the signed chi-squared approach has comparable or better results than diverse density under a varying number of conditions.

The chi-squared approach is very general and can be used for a number of different concept classes beyond those presented here. Any concept class, differentiable or not, can

be used with this approach. This means that other concepts classes can make use of the pruning mechanism described here.

Although the contingency table that we presented relies on the bags having binary labels (e.g., positive and negative), it could be extended to real-valued labels. This is an important task as the real-valued labels can provide additional information to an MI learner (Amar et al., 2001). A related aspect of the diverse density definition is the ability to weigh the evidence in the positive or negative bags differently by separately varying σ_+^2 and σ_-^2. This fits into the signed chi-squared framework in the mechanism for labeling bags.

Acknowledgments

The authors are grateful for comments and helpful discussions from Kiri Wagstaff, Jennifer Neville, and Hannah Blau. This effort is supported by DARPA and AFRL under contract numbers F30602-00-2-0597 and F30602-01-2-0566. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements either expressed or implied, of DARPA, AFRL, or the U.S. Government.

References

Amar, R. A., Dooly, D. R., Goldman, S. A., & Zhang, Q. (2001). Multiple-instance learning of real-valued data. *Proceedings of the Eighteenth International Conference on Machine Learning* (pp. 3–10). Morgan Kaufmann, San Francisco, CA.

Dietterich, T. G., Lathrop, R. H., & Lozano-Perez, T. (1997). Solving the multiple-instance problem with axis-parallel rectangles. *Artificial Intelligence, 89*, 31–71.

Goldman, S., Zhang, Q., Yu, W., & Fritts, J. E. (2002). Content-based image retrieval using multiple-instance learning. *Proceedings of the Nineteenth International Conference on Machine Learning* (pp. 682–689). Morgan Kaufmann, San Francisco, CA.

Maron, O. (1998). *Learning from ambiguity*. Doctoral dissertation, Massachusetts Institute of Technology.

Maron, O., & Lozano-Pérez, T. (1998). A framework for multiple-instance learning. *Advances in Neural Information Processing Systems 10* (pp. 570–576). Cambridge, Massachusetts: MIT Press.

Maron, O., & Ratan, A. L. (1998). Multiple-instance learning for natural scene classification. *Proceedings of the Fifteenth International Conference on Machine Learning* (pp. 341–349). Morgan Kaufmann, San Francisco, CA.

McGovern, A. (2002). *Autonomous discovery of temporal abstractions from interaction with an environment*. Doctoral dissertation, University of Massachusetts Amherst.

McGovern, A., & Barto, A. G. (2001a). Automatic discovery of subgoals in reinforcement learning using diverse density. *Proceedings of the Eighteenth International Conference on Machine Learning* (pp. 361–368). San Francisco, CA: Morgan Kaufmann Publishers.

McGovern, A., & Barto, A. G. (2001b). Linear discriminant diverse density for automatic discovery of subgoals in reinforcement learning. Poster presentation at the Workshop on Hierarchy and Memory in Reinforcement Learning at the 18th International Conference on Machine Learning.

McGovern, A., & Jensen, D. (2003). Identifying predictive structures in relational data using multiple instance learning. *To appear in the Proceedings of the Twentieth International Conference on Machine Learning*.

Oates, T., & Cohen, P. R. (1996). Searching for structure in multiple streams of data. *Proceedings of the Thirteenth International Conference on Machine Learning* (pp. 346–354). Morgan Kauffman.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, California: Morgan Kaufmann Publishers.

Rosenstein, M. T., & Barto, A. G. (2001). Robot weightlifting by direct policy search. *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence* (pp. 839–844). San Francisco: Morgan Kaufmann.

Webb, G. I. (1995). OPUS: An efficient admissible algorithm for unordered search. *Journal of Artificial Intelligence Research, 3*, 431–465.

Zhang, Q., & Goldman, S. A. (2002). EM-DD: An improved multiple-instance learning technique. *Advances in Neural Information Processing Systems 14*. Cambridge, MA: MIT Press.